

# ProGen: Language Modeling for Protein Generation

Year: 2020 | Citations: 312 | Authors: Ali Madani, Bryan McCann, N. Naik, N. Keskar, N. Anand

---

## Abstract

Generative modeling for protein engineering is key to solving fundamental problems in synthetic biology, medicine, and material science. We pose protein engineering as an unsupervised sequence generation problem in order to leverage the exponentially growing set of proteins that lack costly, structural annotations. We train a 1.2B-parameter language model, ProGen, on ~280M protein sequences conditioned on taxonomic and keyword tags such as molecular function and cellular component. This provides ProGen with an unprecedented range of evolutionary sequence diversity and allows it to generate with fine-grained control as demonstrated by metrics based on primary sequence similarity, secondary structure accuracy, and conformational energy.